Detecting Multidimensionality: Which Residual Data-type Works Best?

John Michael Linacre
University of Chicago

ABSTRACT

Factor analysis is a powerful technique for investigating multidimensionality in observational data, but it fails to construct interval measures. Rasch analysis constructs interval measures, but only indirectly flags the presence of multidimensional structures. Simulation studies indicate that, for responses to complete tests, construction of Rasch measures from the observational data, followed by principal components factor analysis of Rasch residuals, provides an effective means of identifying multidimensionality. The most diagnostically useful residual form was found to be the standardized residual. The multidimensional structure of the Functional Independence Measure (FIM^{SM}) is confirmed by means of Rasch analysis followed by factor analysis of standardized residuals.

INTRODUCTION

The Rasch model constructs a one-dimensional measurement system from ordinal data, regardless of the dimensionality of those data. Empirical data are always manifestations of more than one latent dimension. For instance, in observational instruments, the observer's own training level and perspectives influence the observations recorded. In self-administered tests, the ability of the subject to comprehend and follow instructions becomes part of the subject's self-assessment. Consequently the Rasch dimension is a composite based on the conjoint ordering of persons, items and other facets of measurement according to their raw scores (with allowance for incomplete data).

When the data accord exactly with the Rasch model, then all systematic variation within the data is explained by the one dimension. The removal of the implications of this dimension (for both persons and items) from the data leaves behind observation-level residuals with a random normal structure and predictable variance (Wright & Masters, 1982, p. 98). Consequently, the residuals for pairs of items across persons are uncorrelated, a property known as "local independence" (Lazarsfeld, 1958). Since Lazarsfeld introduced the term "local independence" in the context of latent class analysis, he conceptualized all relevant persons to be located at the same point on the variable. In Rasch usage and in this paper, local independence is modeled to hold not just for the classes, i.e., at particular points along the variable, but at every point along the variable. Thus local independence is modeled to hold not just at the class level, but for each person. To verify local independence under Rasch model conditions, for which replication of observations is necessary, coincidence of person locations on the latent variable is achieved by removing the effect of different person measures from the observations (Andrich, 1991).

In practice, however, it is impossible to discern, from the data alone, whether a particular residual is an accidental outcome of a process that accords with the Rasch model, or is produced by

unmodeled dimensions. Indeed, all deviation in the data from the Rasch dimension could be considered symptoms of multidimensionality. Is an unexpected correct answer on a test the one-in-a-thousand occurrence predicted by the Rasch model, or is it a lucky guess? Even a single random lucky guess on a certification test results in data that confounds a competence dimension with a guessing dimension, causing the Rasch dimension to be a composite of the two. Since the certification information in the data overwhelms the guessing information, most users are content to label the test a "certification" test, and the Rasch dimension, a "certification" dimension.

A few unusual responses slightly bias the measures toward the center of the test (Adams and Wright, 1994). They also slightly reduce the statistical validity of the measures of the relevant persons and items (Wright and Stone, 1979, pp. 181-190). When such observations are a cause for concern, they can be identified and diagnosed by examining the patterns of responses by the relevant persons or to the relevant items. Since such detailed examination of all the data is unreasonable, it is useful to perform an initial screening of the data using person- and item-level quality control fit statistics, such as Outfit and Infit (Wright and Stone, 1979, pp. 66-82). Gross non-normality of residuals would be detected at this stage.

A pervasive, but usually less obvious, perturbation of the residuals is symptomatic of the presence of more than one dimension in the data. Extra dimensions may reflect different person response styles or different item content areas. Since, unidimensionality is always provisional, and ultimately utilitarian, the occurrence of multiple dimensions in the data does not necessarily imply substantive multi-dimensionality. Certification tests contain both theory and practice aspects, but the data can express a unidimensional "competency" variable.

Multidimensionality only becomes a real concern when there are response patterns in the data indicating that the data represent two or more dimensions so disparate that it is no longer clear what latent dimension the Rasch dimension operationalizes. A data-set manifests one dimension so long

as it is productive to think of it that way. For educational policy-makers, math is everything from addition to calculus. For cognitive psychologists, the mental processes underlying addition may be very different from those underlying subtraction.

In the extreme, every test item defines its own dimension. For instance, a common one-item test is the question, "What is your age"? In diagnostic testing, each response to each item may indicate a specific course of action. Nevertheless, the inferential goal is to generalize across as many different items as possible that usefully manifest the same variable, such as "patient independence". Utility is defeated, however, when different subsets of such items would lead to different generalizations. In this instance, utility dictates that what was considered to be the "same" variable is, in fact, two (or more) different variables, each leading to different inferences. An example is the Functional Independence Measure, FIM^{SM}. Though originally intended to generalize one dimension of functional independence across a mixture of 18 motor and cognitive items, closer inspection indicated that it would generally be more useful to use the FIM items to construct separate "motor" and "cognitive" measures for each patient (Linacre et al., 1994).

Multidimensionality can also be an artifact of test construction. For instance, including the identical item several times in a certification test produces a subset of responses to those items that have high inter-correlation across persons. These items define their own idiosyncratic local dimension based on that one item. On the other hand, the use of different response mechanisms across items (multiple-choice, open-ended, rating scales) introduces unmodeled variation in the response-level data that can be attributed to a dimension of "item type" (Wilson and Wang, 1995).

Identifying Statistical Multi-Dimensionality

Since the only multidimensionality of real measurement concern is manifested by unmodeled behavior in the data, it is that part of the data that must be examined. After the construction of Rasch measures from the current data (or their imputation from previous data or by theory), an

expected value can be computed for each ordinal observation. The observation residual is the observation less its expectation. It is by looking for patterns among these residuals that relevant multidimensionality can be identified. "Analysis of the fit of data to [local independence] is the statistical device by which data are evaluated for their measurement potential - for their measurement validity" (Wright 1995).

Since there are many ways in which data can depart from the Rasch model (Glas and Verhelst, 1995), it has been suggested that the most blatant departures be investigated first, followed by more subtle ones. Using a comparative example, Linacre (1992) suggested a three stage procedure: (i) remediate systematic contradictions to the Rasch dimension, typically flagged by negative point-biserial correlations; (ii) diagnose idiosyncratic persons and items using local quality-control fit statistics, such as INFIT and OUTFIT; (iii) look for multidimensionality.

It is the residual inter-correlations across items that indicate whether subgroups of items cluster together in a non-homogeneous way, symptomatic of multidimensionality. "The misfit of the Rasch model to a data set can be measured by the size of residual covariances. Unfortunately, some computer programs for fitting the Rasch model do not give any information about this. A choice would be to examine the covariance matrix of the item residuals, not the sizes of the residuals themselves, to see if the items are indeed conditionally uncorrelated, as required by the principle of local independence" (McDonald, 1985, p. 212).

Conditionally correlated item residuals indicate the presence of other measurement dimensions, beyond the primary dimension. This suggests a two-step process. First, identify the other dimensions. Second, decide whether they are of sufficient interest to warrant the construction of separate measures for those dimensions. This paper focuses on the first step, the identification of secondary dimensions. In this endeavor, principal components factor analysis is used to detect structure in the inter-item residual correlation matrix.

The use of factor analysis to identify the primary dimension in data is discussed by Wright (1996) and Smith (1996). In essence, factor analysis aids in the classification of items into potential dimensions, and assists with the partitioning of raw scores according to those dimensions. It does not however, construct linear measures from the data along those dimensions. Consequently, factor scores and loadings have an uncertain sample dependency and analyst-choice-dependent nature that renders their direct use in subsequent analyses precarious.

Choice of Residual Form for Item Correlations

Consider a simple polytomous form of the Rasch model:

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) \equiv B_n - D_i - F_k$$

whereP_{nik} is the probability of being observed in ordered category *k* for person *n* on item *i*, where *k* ranges from 1 to *m*.

$P_{ni(k-1)}$ is the probability of being observed in category *k-1* for person *n* on item *i*,

B_n is the ability of person *n*,

D_i is the difficulty of item *i*, and

F_k is the step difficulty of category *k* relative to category *k-1*.

Each data-point, X_{ni}, is an observed category in the range 0 to *m*, resulting from an interaction between person *n* and item *i*. Corresponding to each X_{ni} is an expected score, E_{ni}, given by

$$E_{ni} = \sum_{k=0}^{m} k\, P_{nik}$$

with model variance of the observed outcome about the expected, V_{ni}, where

$$V_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 \quad P_{nik} = \sum_{k=0}^{m} k^2 P_{nik} - E_{ni}^2$$

This suggests a variety of residuals for investigation regarding inter-item correlation (see Table 1). The raw score residual, Y_{ni}, is the difference between the observed and expected category values and has the range *-m* to *m*. Each standardized residual, Z_{ni}, is normalized by its local modeled standard deviation. These standardized residuals are expected to approximate a N(0,1) distribution (Smith, 1988). The logit residual, L_{ni}, is a first approximation to the measurement discrepancy indicated by the raw score residual. The modeled observation variance, V_{ni}, is the raw-score-to-logit conversion factor (Wright and Masters, 1982, p. 77). The relationship between the three residuals can be complex, and depends on the shape of the item information function, defined by the rating scale structure. For a two-category rating scale, i.e., for dichotomous observations, the relationship is shown in Figure 1.

[Table 1 about here]

[Figure 1 about here]

The choice of which type of residual to employ in the investigation of multidimensionality is not clear cut. A *prima facie* case could be made for each one of them. Since Rasch analysis is a measurement-based approach, investigation of residuals from a measurement-based perspective would appear most productive. This would focus on the logit residuals. On the other hand, since unmodeled patterns in the residuals contradict the measurement framework, the standardized residuals may have more diagnostic power due to their clear statistical properties. The raw score residuals, however, most directly reflect the presence of any other dimensions. Indeed, these last most closely resemble the original raw observations which are widely used in the investigation of multidimensionality (Thurstone, 1932).

A Simulation Study for Two Dimensions

In view of the uncertainty in the choice of residual with which to compute inter-item correlations, a

series of simulation studies was conducted. The purpose of the studies was to discover which form of residual most clearly identified the multidimensional structure underlying the data in straight-forward situations. Principal components analysis, also called the principal-factor method, was chosen for this investigation because of its "rigorous mathematical basis" (Harman, 1960, p. 154). Substitution of common-factor methods in these simulation studies (not reported here) was found to lead to the same conclusions. The simulation studies employ dichotomous items, but the utility of their result is illustrated with a polytomous empirical data set.

For the first study, a sample of 1190 persons was generated. Each person was assigned two orthogonal abilities: a "math" ability randomly from an N(0,1.5) logit distribution, and a "reading" ability randomly from a distribution with the same shape. The two abilities were assigned independently, producing orthogonality. A two-dimensional test was then posited containing 3 types of dichotomous items: (i) 100 "math" items uniformly distributed in difficulty over -2 to +2 logits; (ii) 25 "reading" items uniformly distributed over -2 to +2 logits; (iii) 50 "word problem" items (conceptually combine reading and math) uniformly distributed over -2 to +2 logits.

Dichotomous observational data were generated for each person. For the math items, the math ability was used. For the reading items, the reading ability was used. For the word problem items, the *lower* of each person's math and reading abilities was used.

Rasch analysis of this observational data was performed. One measure was estimated for each person across all items and one difficulty for each item across all persons using the BIGSTEPS Rasch analysis program (Wright and Linacre, 1997). Based on these estimated measures, expected observations were obtained and the three score residuals calculated.

Because there are more math than reading items, the primary "Rasch" dimension is expected to be dominated by the math items. The reading items should give the strongest indication of a second

dimension. The word problems should cluster halfway between the math and reading items. Smith and Miao (1994) reported that the ratio of 4 items on one dimension to 1 item on another generally produces a dimensional structure that can be identified directly by principal components analysis of the observations themselves. Accordingly, this was done.

[Figure 2 about here]

Figure 2 shows a plot of the loadings of the first principal component (unrotated) in the simulated data against the Rasch item difficulties estimated from that same data. (Computations were performed by the author using proprietary software which had been validated against standard data sets). The item difficulties fall mainly within their simulated range of -2 to +2 logits. The 0.5 logit increase in the difficulty of the "W" items (word problems) relative to their generators is due to the choice of the lower of math and reading ability in generating the observations. This choice has had the expected effect of making the estimated items appear more difficult than the generating items.

The loadings on the first principal factor in the observations stratify the items by type: M for math items, W for word problems and R for reading items. The math items show the highest loading on the first factor, the reading items the least, as expected. The effect of item difficulty level is secondary, but the convex form of the "M" distribution indicates that extreme item easiness or difficulty attenuates the loading on the first factor. The non-linearity of raw scores is distorting the factor structure. Consequently, the vertical difference between the lowest M and highest W is small, meaning that the stratification, which is obvious in the plot, would be less striking in a table of factor loadings.

[Figure 3 about here]

Figure 3 is based on the raw residuals. The factor loadings for the first principal component in the

item residual correlations are plotted against item difficulties. The first (Rasch) dimension has been explicitly removed. The highest loadings on this second, residual, dimension are now obtained by the reading items. (Since factor direction is arbitrary, the largest factor loading is shown as positive in this study). The fact that the Rasch dimension is a compromise between the math and reading items is confirmed by the negative, rather than zero, loadings of the math items.

The raw residuals produce a better stratified and less curved plot than the original observations. This could have been expected because the data were simulated to fit the Rasch model. Nevertheless, it is encouraging that introducing another orthogonal dimension into the data has not invalidated a Rasch-based dimensional structure.

[Figure 4 about here]

Figure 4 is based on the standardized residuals. With these data, the differences between the standardized and raw residual plots are barely distinguishable by eye.

[Figure 5 about here]

Figure 5 employs the logit residuals. This plots shows attenuated loadings on the extreme items, clouding the nature of the dimensionality in the simulated data. Nevertheless, this Figure remains clearer than that based on the observations themselves, Figure 2.

These simulations of dichotomous observations suggest that none of these four approaches would be misleading, but that raw and standardized residuals give the clearest results.

Simulation Study: Correlated Dimensions

A more subtle form of multidimensionality is that of correlated dimensions. As a trainee advances

through a course of study, both knowledge of theory and practical skills tend to improve, but not exactly in step. This can lead to the trainee having a "knowledge" ability and a "skill" ability. Across a sample of trainees at different stages of their training these abilities will be correlated, but different. A test consisting of both knowledge and skill items will probe both abilities, and the reported trainee measure will be a composite of the two abilities. Analysis of residuals can alert the analyst that this has occurred.

In the second simulation, a sample of 1000 persons was generated. Each person was assigned two abilities: an "X" ability randomly from an N(0,1) logit distribution, and a "O" ability randomly from a distribution with the same shape, but such that the X and O abilities have a 0.9 correlation across the sample. Responses by this sample to a test of 50 X-type and 50 O-type items were simulated, such that each person is modeled to respond to each item type with the corresponding ability, e.g., responses to X items are with X abilities. For each item type, the item difficulties were uniformly distributed from -2.0 to +2.0 logits.

The inter-ability correlation of 0.9 was set high so that neither principal components factor analysis of the observations nor item-level OUTFIT statistics would be expected to detect the dimensional nature of the items successfully (Smith and Miao, 1994). As a further complication, the mean ability of the sample was set at the center of the test, removing any skewing of the observation variance.

[Figure 6 about here]

Principal components factor analysis of inter-item correlations was performed. Figure 6 shows the loadings on the *second* factor for these simulated observations. This factor is generally successful in discriminating X and O-type items. The most displaced X and O items are indicated with arrows.

[Figure 7 about here]

Figure 7 shows the loadings on the first factor for the logit residuals. This approach is less successful in discriminating X and O items. In particular, the most displaced X item, at the bottom of the plot, is indicated to be more strongly O-type than nearly all O items.

[Figure 8 about here]

Figure 8 shows the loadings on the first factor for the raw residuals. This approach is more successful. Only one O item and one X item are noticeably displaced.

[Figure 9 about here]

Figure 9 shows the loadings on the first factor for the standardized residuals. This approach is the most successful. Only one X item is noticeably displaced.

In similar simulations, not reported here, but with lower inter-dimensional correlations and different sample-test targeting, this pattern continued. The logit residuals were the least successful at discriminating X and O type items. Factor analysis of the observations themselves was more successful in discriminating item types, but the raw and standardized residuals were most successful and about equally effective.

An Example Application

In order to verify the effectiveness of principal components factor analysis of residuals, Rasch analysis was performed on a random sample of 6,144 FIM^{SM} records (from the UDS database,

courtesy of Carl V. Granger). Only data collected at the *admission* time point were analyzed. Figure 10 plots the loadings of the first factor in the standardized residuals against the logit calibrations of the 18 FIM items. This Figure immediately signals the divergence of the five cognitively-oriented items (top of the Figure) from the thirteen motor-oriented items. This same divergence was reported in Linacre et al. (1994), but only after a tortuous analysis of admission and discharge data. For these FIM data, the raw residual plot was almost identical to Figure 10, but with slightly less range to the loadings. Both identify the opposite poles of the factor to be "memory" and "toilet transfer". Analyses of the logit residuals and the original observations each generated a minor factor that corresponded to the cognitive-motor contrast, but with different orderings of the items at each end of the factor. For the raw observations, the extremes are "comprehension" vs. "stairs". For the logit residuals, "bathing" vs. "problem solving". Thus, though the standardized residuals provided the distinct solution, the clinical implications of the factor structure might direct the analyst to favor use of a different residual for this analysis.

[Figure 10 about here]

Once divergence within an item pool has been identified, the next step is to evaluate its impact on measurement. For the FIM, this is investigated by measuring the sample, first on the variable defined by the five cognitive items, then on that defined by the thirteen motor items. When the differences between the resulting pairs of measures have clinical implications, e.g., when one measure indicates normal functioning and the other dysfunction, the multidimensionality of the original instrument is resolved by setting up two measurement systems. This is the case for many applications of the FIM. When differences between the pairs of measures have no implications for practice, then the multidimensionality is treated as an unwanted, but inevitable, source of the noise within the data, slightly lowering the quality of the one measurement system.

Conclusion

For complete tests, principal components factor analysis of either the observations themselves or the various residual formulations successfully reflects the multidimensional structures simulated here. Though these simulated structures are more clear-cut than those hypothesized to exist in empirical data, the essential features are likely to encompass the same structures: items of varying dimensionality and persons with multiple, but correlated, abilities. A word of caution: empirical data often incorporate departures from the Rasch model that would distort the distribution of the residuals, including miskeyed items, data entry errors and response sets. It is recommended that these issues be addressed prior to factor analysis.

Overall, standardized residuals provided the most decisive analysis, but their advantage over raw residuals was slight. Logit residuals were less informative.

Factor analysis of the observations themselves was also informative of the factor structure, but with the huge impediment that it does not construct linear measures for even one of its many factorial dimensions. Further, it requires the analyst to determine which factor reflects the predominant measurement system, and which the multidimensionality. Factor rotation or other factor methods may clarify this, but they can also confuse the factor structure further (Ferguson, 1941).

In this study, Rasch analysis followed by factor analysis of residuals was always more effective at both constructing measures and identifying multidimensionality than direct factor analysis of the original response-level data.

REFERENCES

Adams, R. J., and Wright, B. D. (1994) When does misfit make a difference? In M. Wilson (Ed.) *Objective Measurement: Theory into Practice, Vol. 2,* 244-270. Norwood, NJ: Ablex Publishing Corporation.

Andrich, D. A. (1991) Local independence, latent class analysis and Rasch. *Rasch Measurement Transactions*, *5, 3*, 160-161.

Ferguson, G. A. (1941) The factorial interpretation of test difficulty. *Psychometrika 6, 5,* 323-329

Glas, C. A. W., and Verhelst, N. D. (1995) Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (Eds.) *Rasch Models: Foundations, Developments and Recent Applications*, 69-96. New York: Springer Verlag.

Harman, H. H. (1960) *Modern Factor Analysis.* Chicago: University of Chicago Press.

Lazarsfeld, P. F. (1958) Latent structure analysis. In S. Koch (Ed.) *Psychology: A study of science. Vol. III,* 476-543. New York: McGraw-Hill.

Linacre, J. M. (1992) Prioritizing misfit indicators. *Rasch Measurement Transactions 9, 2*, 422-423.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., and Hamilton, B. D. (1994) The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation 75, 2*, 127-132.

McDonald, R. P. (1985) *Factor Analysis and Related Methods*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, R. M., and Miao, C. Y. (1994) Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.) *Objective Measurement: Theory into Practice. Vol. 2,* 316-327. Norwood, NJ: Ablex Publishing Corporation.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657-667.

Smith, R. M. (1996) A Comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling 3, 1,* 25-40.

Thurstone, L. L. (1932) *The Theory of Multiple Factors.* Ann Arbor, Michigan: Edwards Brothers.

Wilson, M., and Wang, W.-C. (1995) Complex composites: issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19, 1*, 51-71.

Wright, B. D. (1995) Scores, reliabilities and assumptions. *Rasch Measurement Transactions, 5, 3*, 157-158.

Wright, B. D. (1996) Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3, 1,* 3-24.

Wright, B. D., and Linacre, J. M. (1997) *BIGSTEPS Rasch analysis computer program*. Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982) *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B.D., and Stone, M. H. (1979) *Best Test Design*. Chicago: MESA Press.

Acknowledgement:

Suggestions made by Mark Wilson and two anonymous referees have improved this paper.

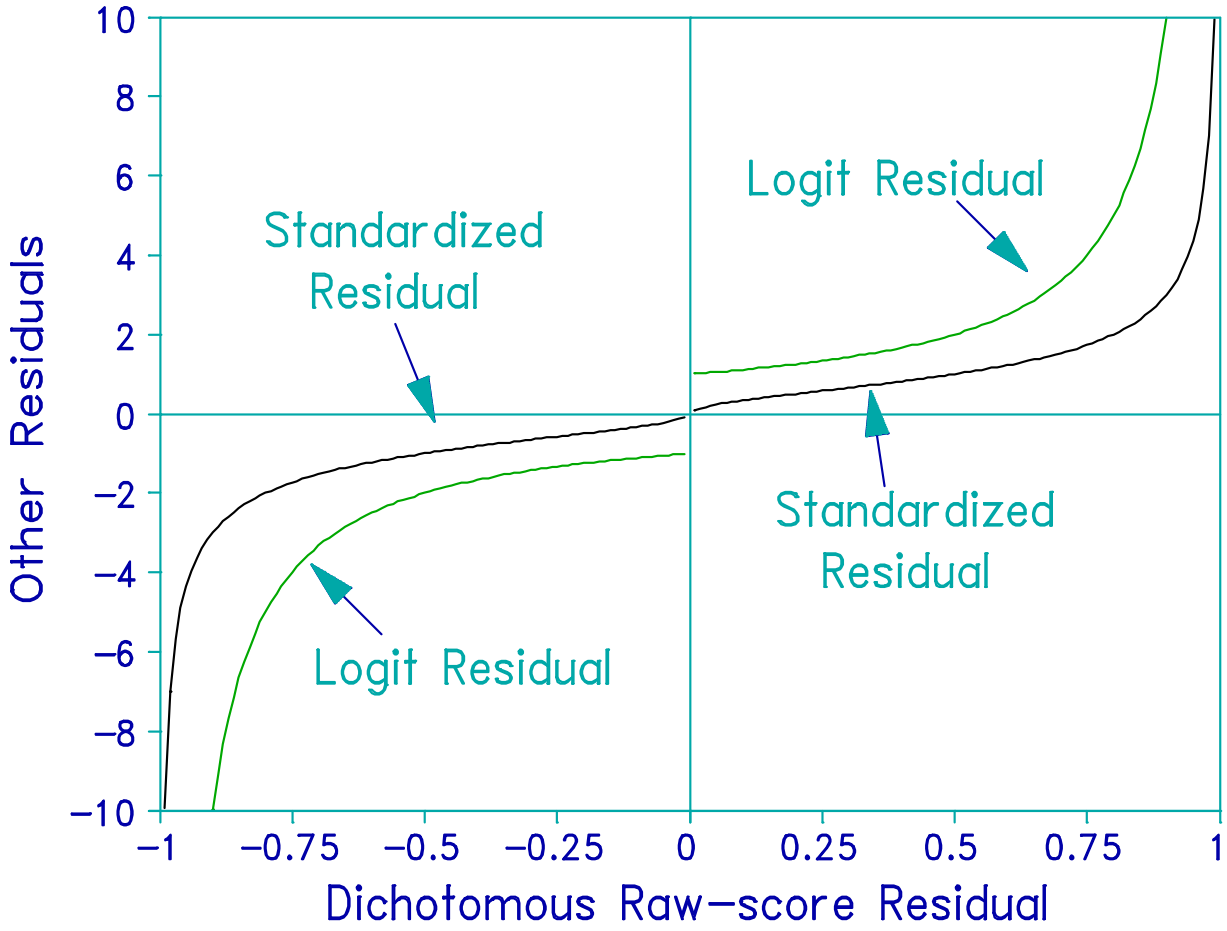

Figure 1. Relationship between residuals to dichotomous observations.

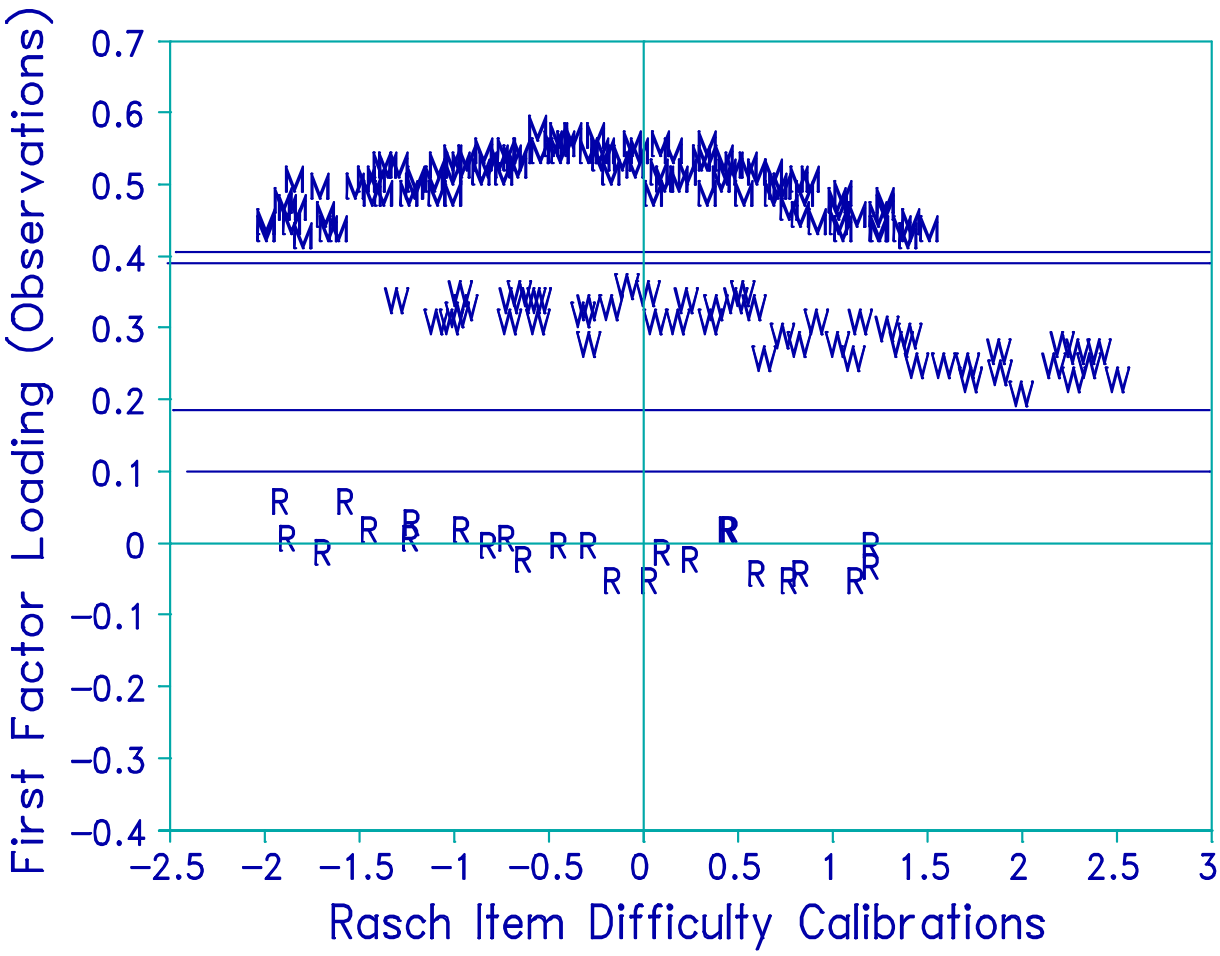

Figure 2. Plot of first principal component of item observations against Rasch item difficulty calibrations.

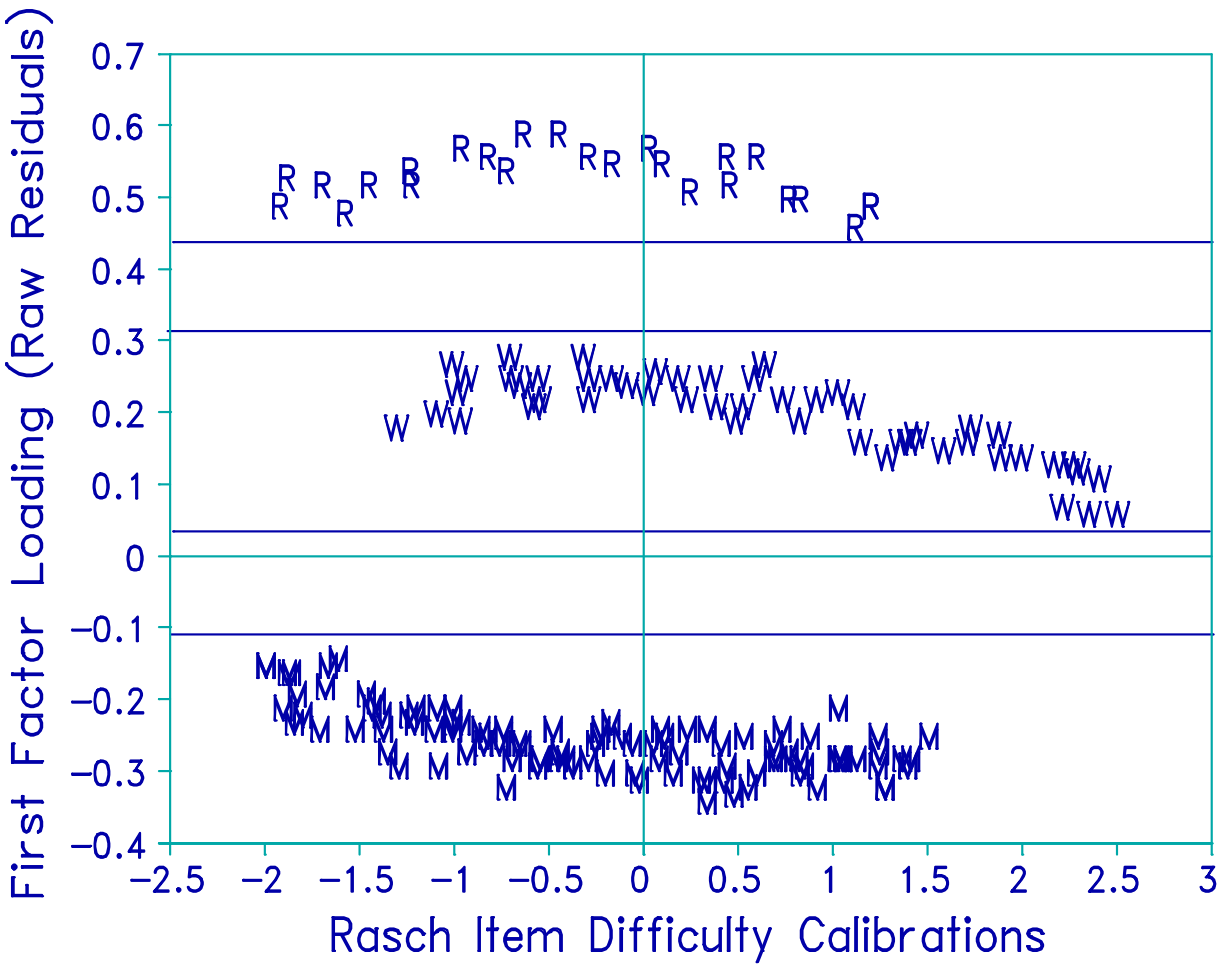

Figure 3. Loadings on first principal component in raw residual correlations against Rasch item calibrations.

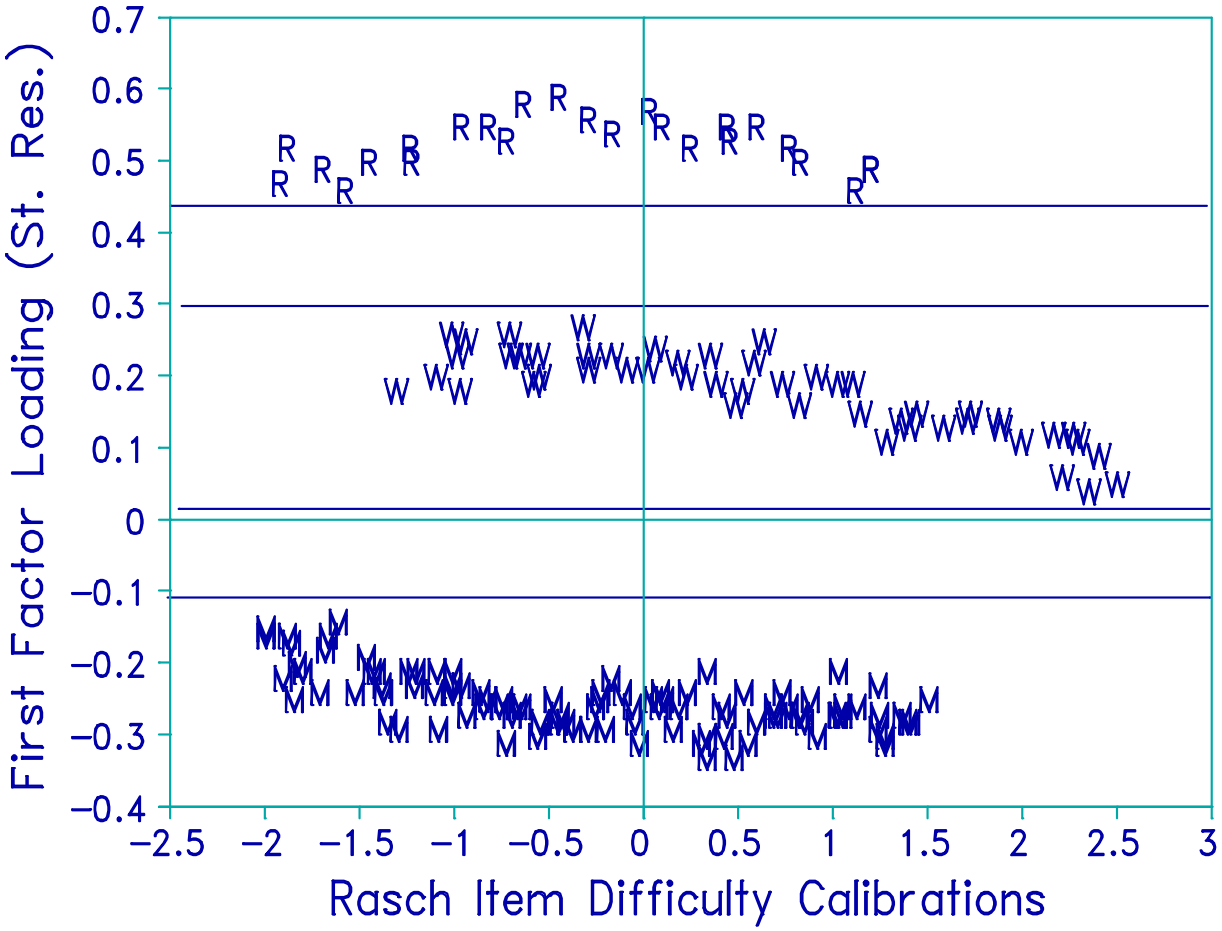

Figure 4. Loadings on first principal competent in standardized residual correlations against Rasch item calibrations.

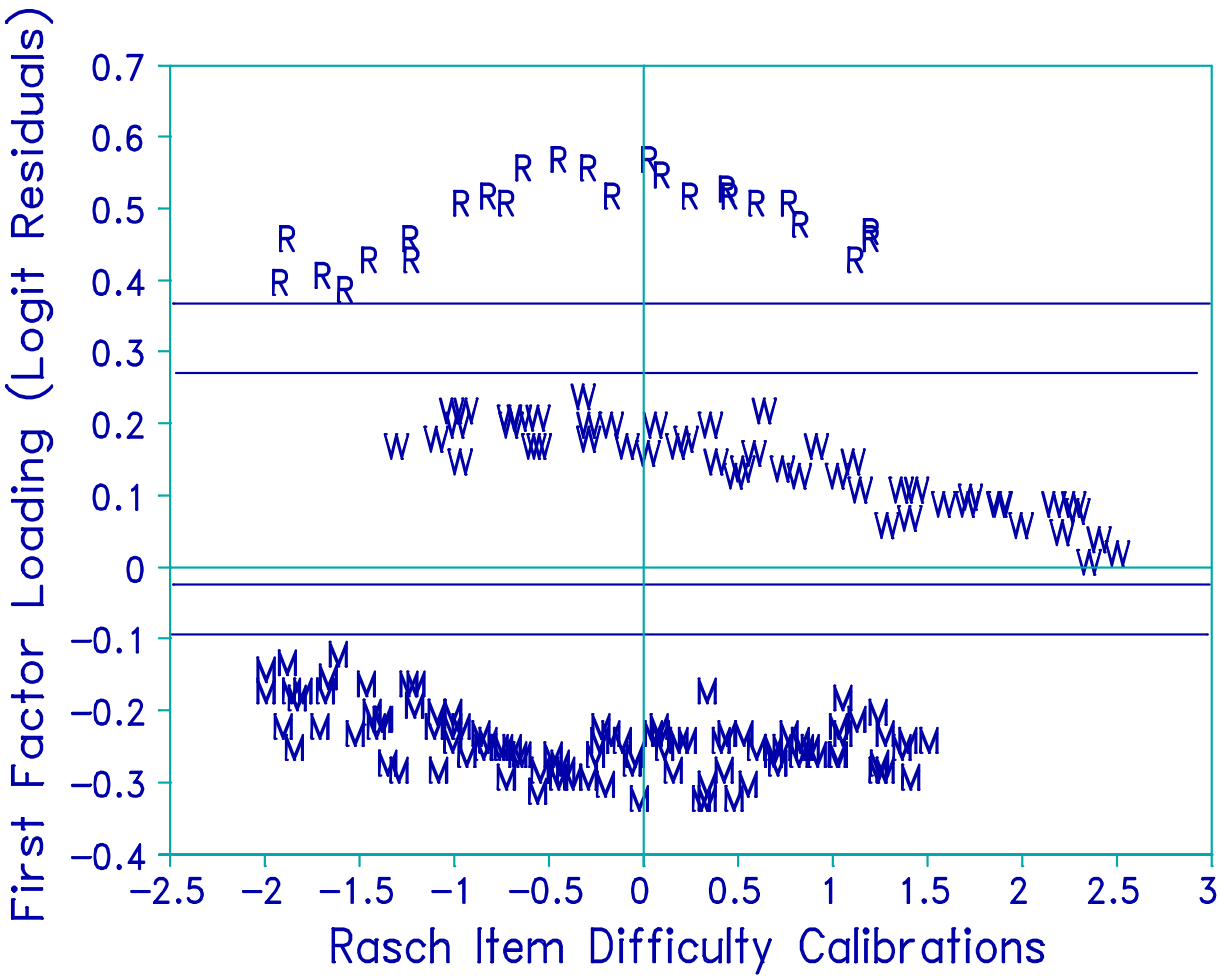

Figure 5. Loadings on first principal component in logit residual correlations against Rasch item calibrations.

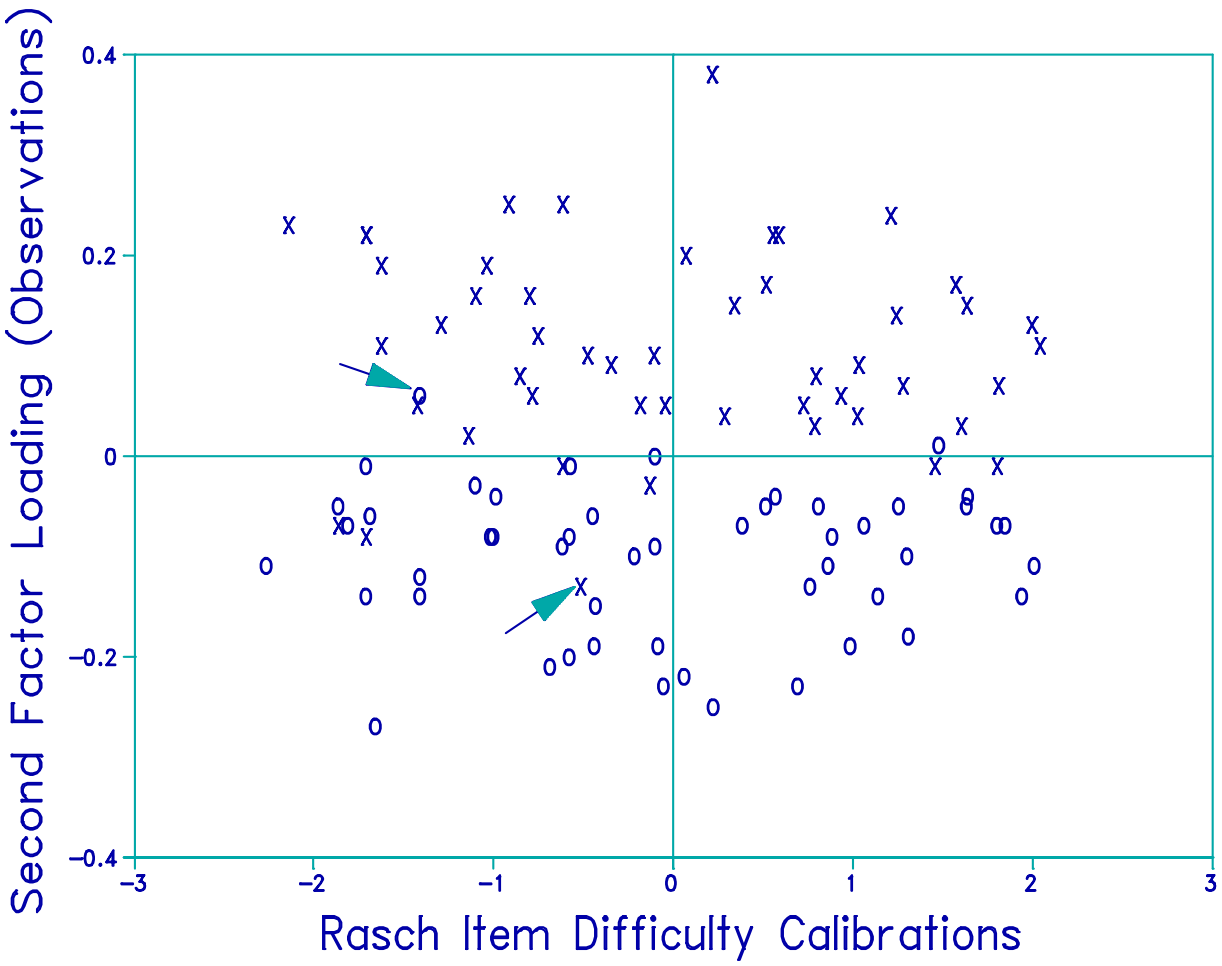

Figure 6. Loadings on the second factor for observations with the correlated multidimensional data.

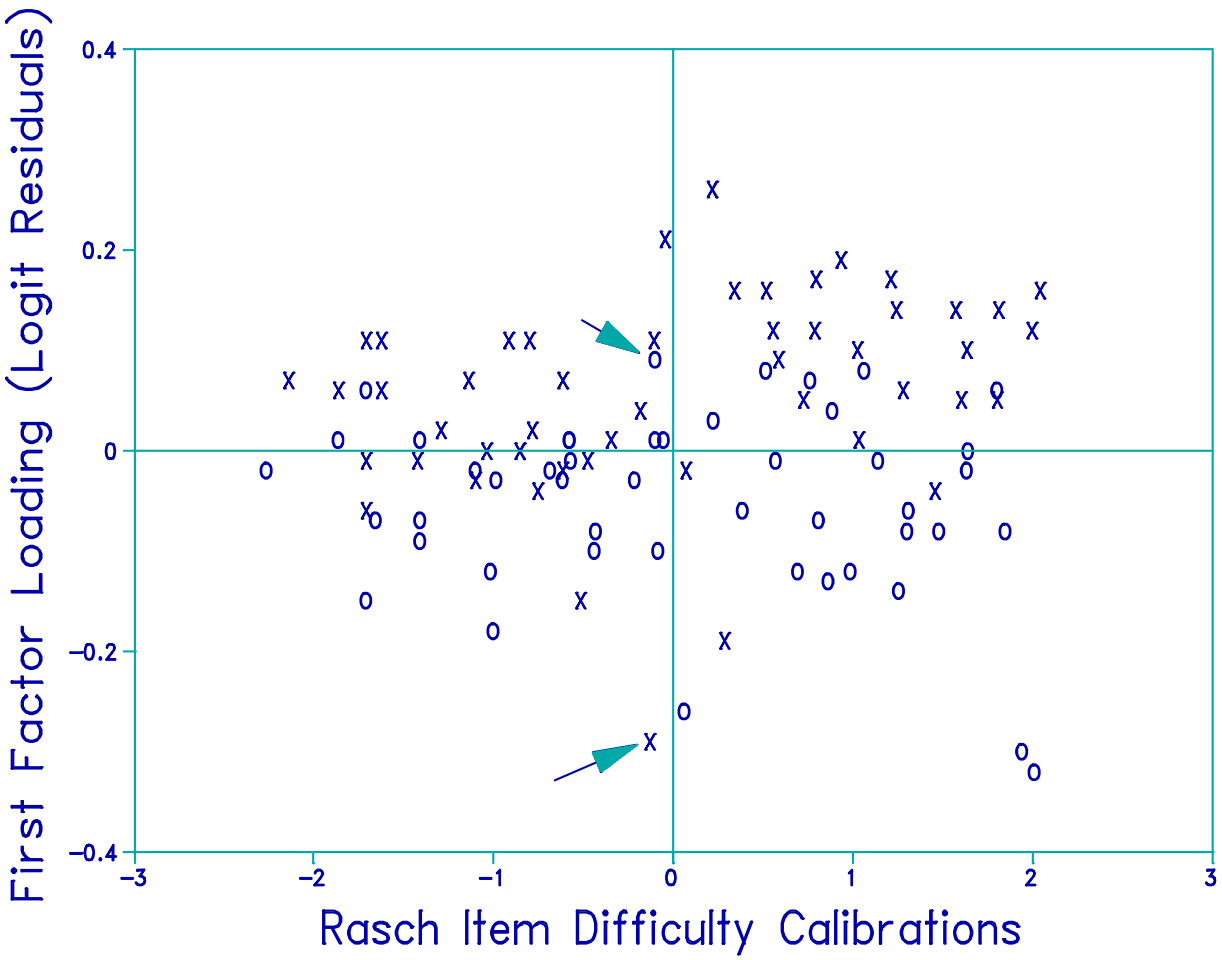

Figure 7. Loadings on the first factor for logit residuals with the correlated multidimensional data.

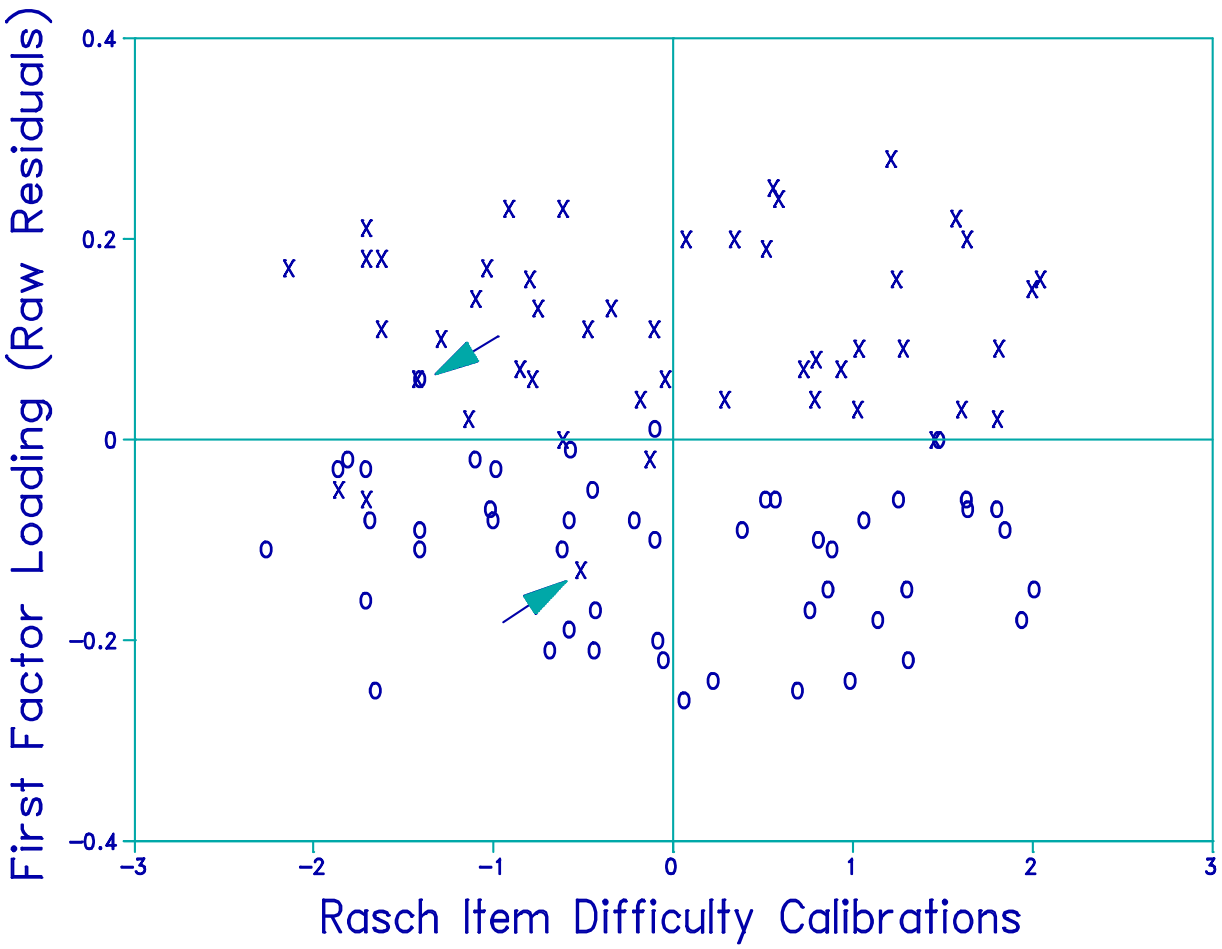

Figure 8. Loadings on the first factor for raw residuals with the correlated multidimensional data.

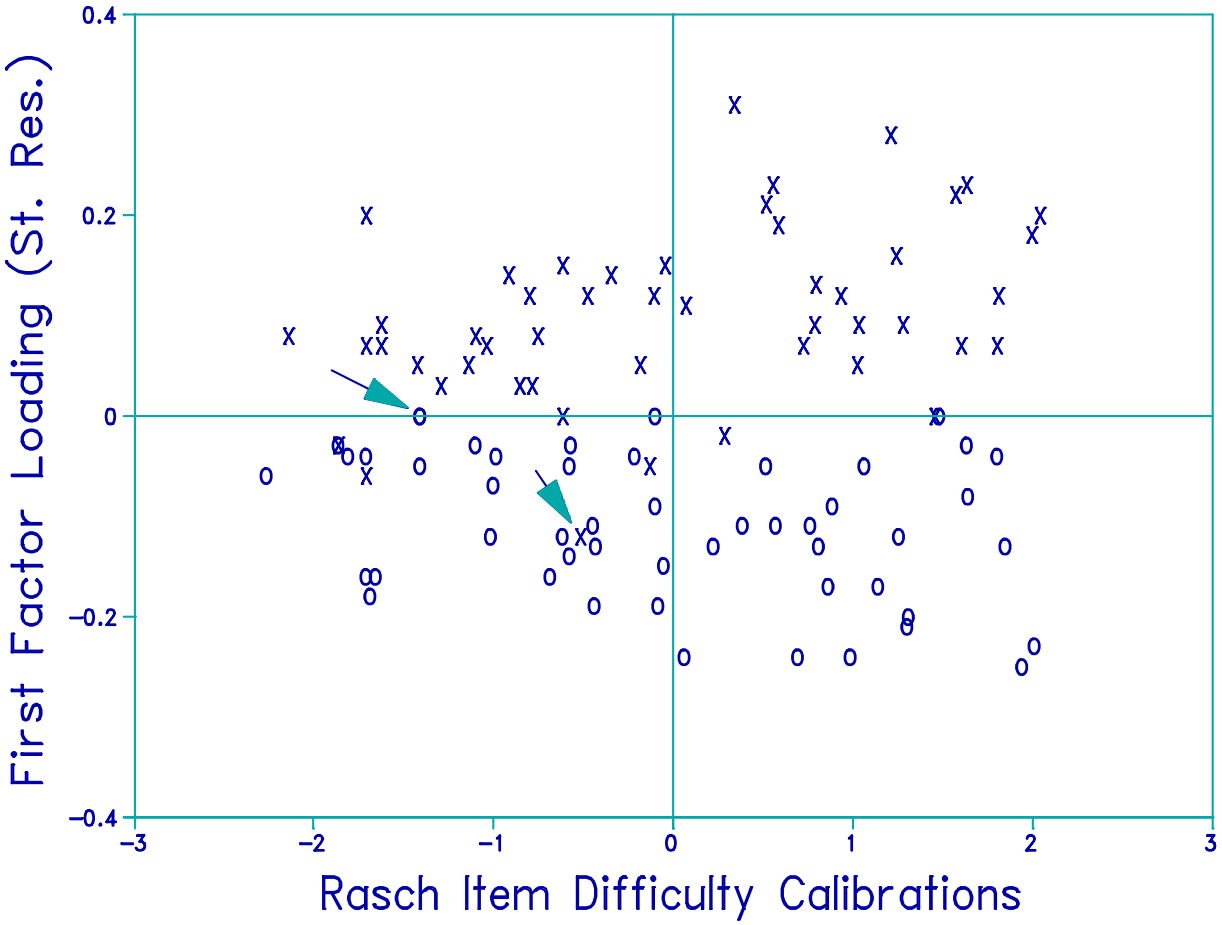

Figure 9. Loadings on the first factor for standardized residuals with the correlated multidimensional data.

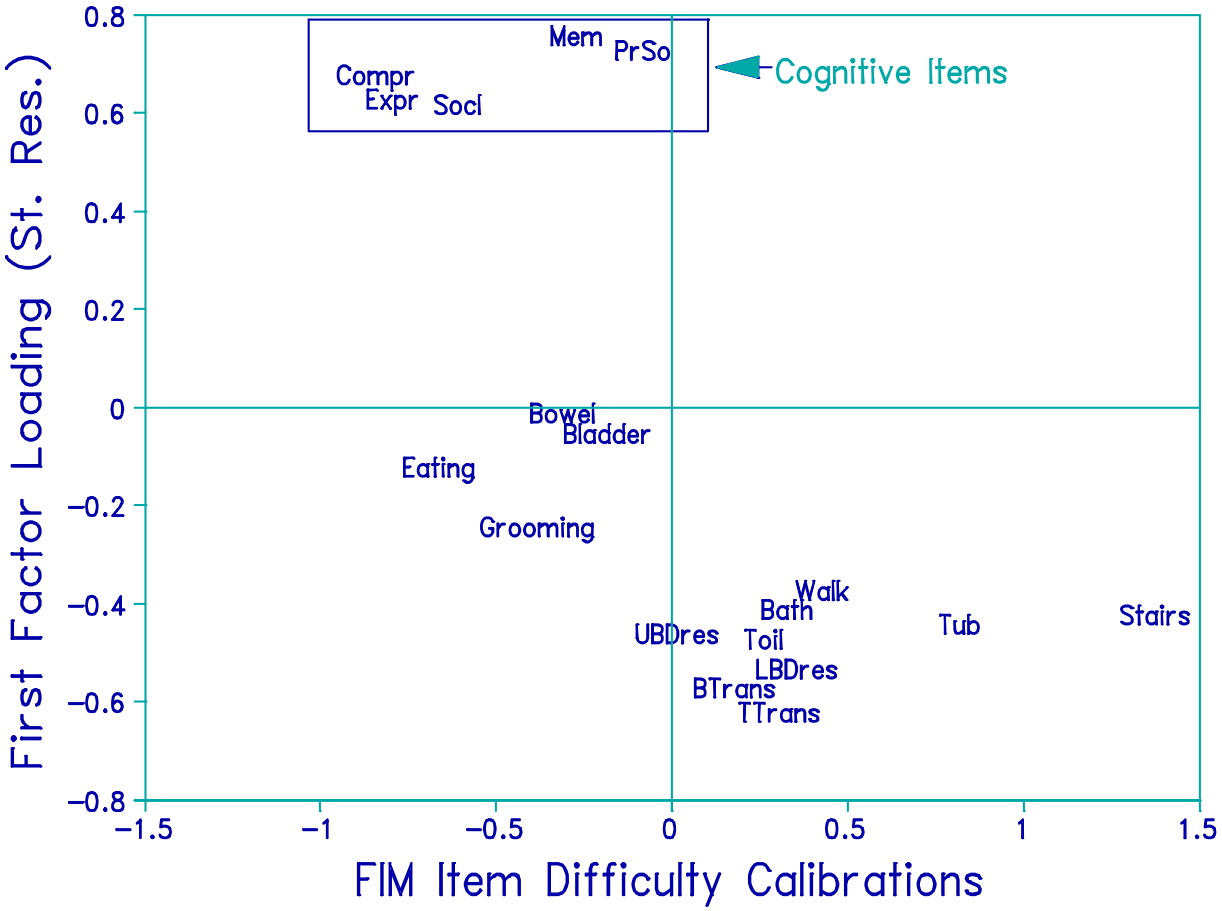

Figure 10. Loadings on the first factor for standardized residuals with the FIM data.

Table 1

Observation Residuals

Residual type	Mathematical expression
Raw score residual	$Y_{ni} = X_{ni} - E_{ni}$
Standardized residual	$Z_{ni} = (X_{ni} - E_{ni})/\sqrt{(V_{ni})}$
Logit residual	$L_{ni} = (X_{ni} - E_{ni})/V_{ni}$